CENTER FOR
Brains
Minds+
Machines

**CBMM Memo No. 103** **July 31, 2020**

# Stable Foundations for Learning: a framework for learning theory (in both the classical and modern regime).

**Tomaso Poggio**[1]

[1]Center for Brains, Minds, and Machines, MIT

## Abstract

I consider here the class of supervised learning algorithms known as Empirical Risk Minimization (ERM). The classical theory by Vapnik and others characterize *universal consistency* of ERM in the *classical regime* in which the architecture of the learning network is fixed and $n$, the number of training examples, goes to infinity. We do not have a similar general theory for the *modern regime* of interpolating regressors and overparamerized deep networks, in which $d > n$ as $n$ goes to infinity.

In this note I propose the outline of such a theory based on the specific notion of *stability* of the learning algorithm with respect to perturbations of the training set. The theory suggests that for interpolating regressors and separating classifiers (either kernel machines or deep RELU networks)

1. minimizing cross-validation leave-one-out stability minimizes the expected error;
2. minimum norm solutions are the most stable solutions;
3. GD algorithms are biased towards a minimum norm solution both for kernel machines and deep classifiers (under the exponential and the square loss).

The hope is that this approach may lead to a theory encompassing both the modern regime and the classical one.

**This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.**

# Stable Foundations for Learning: a framework for learning theory (in both the classical and modern regime).

Tomaso Poggio

July 31, 2020

**Abstract**

I consider here the class of supervised learning algorithms known as Empirical Risk Minimization (ERM). The classical theory by Vapnik and others characterize *universal consistency* of ERM in the *classical regime* in which the architecture of the learning network is fixed and $n$, the number of training examples, goes to infinity. We do not have a similar general theory for the *modern regime* of interpolating regressors and overparamerized deep networks, in which $d > n$ as $n$ goes to infinity.

In this note I propose the outline of such a theory based on the specific notion of $\mathrm{CV}_{loo}$ *stability* of the learning algorithm with respect to perturbations of the training set. The theory suggests that for interpolating regressors and separating classifiers (either kernel machines or deep RELU networks)

1. minimizing cross-validation leave-one-out stability minimizes the expected error;
2. minimum norm solutions are the most stable solutions;
3. GD algorithms are biased towards a minimum norm solution both for kernel machines and deep classifiers (under the exponential and the square loss).

The hope is that this approach may lead to a theory encompassing both the modern regime and the classical one.

## 1 Foundations of Learning Theory

Developing theoretical foundations for learning is a key step towards understanding intelligence. Supervised learning is a paradigm in which natural or artificial networks learn a functional relationship from a set of $n$ input-output training examples. A main challenge for the theory is to determine conditions under which a learning algorithm will be able to predict well on new inputs after training on a finite training set. What should be optimized in ERM to minimize the expected error and, for $n \to \infty$, to achieve consistency? Ideally, we would like to have theorems spelling out, for instance, that consisteny depends on constraining appropriately the hypothesis space.

Indeed a milestone in classical learning theory was to formally show that appropriately restricting the hypothesis space – that is the space of functions represented by the networks – ensures consistency (and generalization) of ERM. The classical theory assumes that the hypothesis space is fixed while the number of training data $n$ increases to infinity. Its basic results thus characterize the "classical" regime of $n > d$, where $d$ is the number of parameters to be learned. The classical theory, however, cannot deal with what we call the "modern" regime, in which the network remains overparametrized ($n < d$) when $n$ grows. In this case the hypothesis space is not fixed.

In trying to develop a theory that can deal with the classical *and* the modern regime, it seems natural to abandon the idea of the hypotehsis space as the object of interest and focus instead on properties of the algorithms. Twenty years ago, while trying to formulate principles of learning beyond ERM (and beyond the use of measures of complexity such as VC dimension, covering numbers and Rademacher numbers), we noted [1] that any supervised learning algorithm is a map $L$ from data sets to hypothesis functions. For a general theory, we asked: *what property must the learning map L have for good generalization error?* The answer was that LOO stability (see [1]) together with $\text{CV}_{loo}$ stability of the algorithm, both going to zero for $n \to \infty$ is sufficient for generalization for any supervised algorithm; $\text{CV}_{loo}$ stability alone is necessary and sufficient for generalization and consistency of ERM. At the time, the surprising connection between stability and predictivity promised a new framework for the foundations of learning theory (see also [2, 3]).

In this paper I sketch how this old proposal may become a learning theory encompassing both the classical and the modern regime for ERM (extensions beyond ERM seem natural but I leave them to future work). I provide several arguments about why low expected error should correspond to stable gradient descent algorithms in the moder n regime. In particular, an interpolating algorithm that minimizes a bound on $\text{CV}_{loo}$ stability should minimize the expected error. Stability optimization may thus provide a unifying principle that could explain, among other properties, the predictivity of deep networks as well as the double descent curve found recently in several learning techniques including kernel machines[1].

### 1.1 Classical Regime

In the classical setting, a key property of a learning algorithm is *generalization*: the empirical error must converge to the expected error when the number of examples $n$ increases to infinity, while the class of functions $\mathcal{H}$, called the *hypothesis space*, is kept fixed. An algorithm that guarantees good generalization will predict well, if its empirical error on the training set is small. Empirical risk minimization (ERM) on $\mathcal{H}$ represents perhaps the most natural class of learning algorithms: the algorithm selects a funcion $f \in \mathcal{H}$ that minimizes the empirical error – as measured on the training set.

One of the main achievements of the classical theory was a complete characterization of the

[1] One may argue that from the point of view of this proposal, the main role of Tikhonov regularization may be to deal with the pathological situation of $d = n$, since asymptotically the inverse of the kernel does not exist if $\lambda = 0$. Of course, presence of noise (significant SNR) has the effect of requiring regularization also for cases close to $d = n$.

necessary and sufficient conditions for generalization of ERM, and for its *consistency* (consistency requires asymptotic convergence of the expected risk to the minimum risk achievable by functions in $\mathcal{H}$; for ERM, generalization is equivalent to consistency). It turns out that consistency of ERM is equivalent to a precise property of the hypothesis space: $\mathcal{H}$ has to be a *uniform Glivenko-Cantelli (uGC)* class of functions (spaces of indicator functions with finite VC dimension are a special case) of uGC .

Our later work [1] showed that an apparently separate requirement – the well-posedness of ERM – is in fact equivalent to consistency of ERM. Well-posedness usually means *existence, uniqueness and stability* of the solution. The critical condition is stability of the solution. Stability is equivalent to some notion of continuity of the learning map (induced by ERM) that maps training sets into the space of solutions, eg $L : Z^n \rightarrow \mathcal{H}$. We recall the definition of *leave-one-out cross-validation (in short, $CV_{loo}$) stability under the distribution $P_S$*:

$$\forall i \in \{1, \ldots, n\} \;\; P_S \left\{ |V(f_S, z_i) - V(f_{S^i}, z_i)| \leq \beta_{CV}^P \right\} \geq 1 - \delta_{CV}, \tag{1}$$

where $V(f, z)$ is a loss function that is Lipschitz and bounded for the range of its arguments and $z = ((x, y)$. $\mathrm{CV}_{loo}$ stability of an algorithm measures the difference between the errors at a point $z_i$ when it is in the training set $S$ of $f_S$ wrt when is not.

We proved [2] that *For ERM, $CV_{loo}$ stability with $\beta_{CV}^P$ and $\delta_{CV}$ in Equation 1 converging to zero for $n \rightarrow \infty$ guarantees, if valid for all $P$, generalization and consistency (and is in fact equivalent to them).*

Notice that $\mathrm{CV}_{loo}$ stability is a weaker requirement than the uniform stability of Bousquet and Elissef which is sufficient but not necessary for consistency of ERM in the classical regime. Of course uniform stability implies $\mathrm{CV}_{loo}$ stability.

### 1.2 Modern Regime

Recently, a different regime has been characterized, first in neural networks [4] and then in linear and kernel regression, mainly because of the pioneering work by Belkin ([5], see also [6] and [7, 8, 5, 9, 10, 11, 12]). In this modern regime, both $n$ (the number of training data) and $d$ (the number of parameters) grow to infinity with $\frac{n}{d}$ constant. If $d \geq n$ there may be exact fitting of the training set and the generalization gap does not go to zero. The classical approach – based on the analysis of the hypothesis space to infer asymptotic generalization and then consistency – cannot be used because there is no fixed hypothesis space. However, the notion of stability, which refers to the algorithm and not the hypothesis space, is not affected by this problem. Since in the "classical" regime of fixed hypothesis space and $n \rightarrow \infty$, stability is important, I expect that a similar notion of stability may work in the "modern" high dimensional regime of $\frac{n}{d} < 1$.

The conjecture discussed in this paper is that *in both cases, stability remains the key requirement for predictivity.* Maximum stability – that is minimum $\beta_{CV}^P$ – is usually guaranteed during minimization of the empirical loss (that is by ERM) by complexity control under the form of regularization (possibly vanishing, as in the definition of the pseudoinverse or as implicitely provided by iterative gradient descent [13]). As I said earlier, the notion of $\mathrm{CV}_{loo}$ stability turns

out to be necessary and sufficient for distribution independent generalization and consistency in the classical framework of ERM with a fixed hypothesis space [2, 1]. In the modern regime, when the empirical error is zero, the definition of $\text{CV}_{loo}$ stability seems closely related to the definition of the expected error for interpolating algorithms (under specific data distributions). It is thus natural to conjecture that *minimization of stability*, in a distribution dependent way, is for ERM a sufficient condition across the classical and the modern regime for minimizing expected error. In the next section I will show that $\text{CV}_{loo}$ stability is almost equivalent in expectation to the expected error for interpolating regressors or classifiers. Then I will discuss the separate conjecture that optimizing $\text{CV}_{loo}$ stability for overparametrized networks is equivalent to selecting minimum norm solutions.

## 2 Stability and Expected error

Let us recall the definition *in expectation of leave-one-out cross-validation (in short, $CV_{loo}$) stability under the distribution $P_S$*:

$$\forall i \in \{1, \dots, n\} \;\; E_S \left| V(f_S, z_i) - V(f_{S^i}, z_i) \right| = \beta_{CV}, \tag{2}$$

where $V(f, z)$ is a loss function that is Lipschitz and bounded for the range of its arguments and $z = ((x, y)$. $\text{CV}_{loo}$ stability of an algorithm measures the difference between the errors at a point $z_i$ when it is in the training set $S$ of $f_S$ wrt when is not.

We want now to consider the case – typical for overparametrized models – of interpolating regressors or separating classifiers, that is the case in which the regressors or classifiers can usually satisfiy $V(f_S, z_i) = 0$, that is they fit the training data under the appropriate loss function (e.g. square loss or classification loss, for instance the function $c$ of [14]). The idea is that then the first term in Equation 2 is negligeable for specific distributions of the data and $\text{CV}_{loo}$ stability becomes essentially equal, in expectation, to the expected loss. This intuition, however, needs to be made rigorous.

To do so, I use the following positivity property of exact ERM [2]

$$V(f_{S^i}, z_i) - V(f_S, z_i) \geq 0. \tag{3}$$

Then $E_S \left| V(f_S, z_i) - V(f_{S^i}, z_i) \right| = E_S[V(f_{S^i}, z_i)] - E_S[V(f_S, z_i)]$
Thus

$$\forall i \in \{1, \dots, n\} \;\; E_S \left| V(f_S, z_i) - V(f_{S^i}, z_i) \right| = E_S I[f_{S^i}] - E_S I_S[f_S] \tag{4}$$

where $I(f_S)$ is the expected error of $f_{S^i}$ and $I_S[f_S]$ is the empirical error of $f_S$. Under specific assumpions on the algoritm and the distribution $P_S$, the term $E_S I_S[f_S]$ can be negligeable, as we will see later. In these cases, $\text{CV}_{loo}$ stability is indeed equal to $I[f_{S^i}]$. In turn for ERM $I[f_{S^i}]$ converges in probability to $I[f_S]$ for $n \rightarrow$
$infty$. As an example of the $I[f_{S^i}]$ term, consider the case in which $V$ is the square loss and

$f_{S^i}(z_i) = W_{S^i} x_i$. Then

$$V(f_{S^i}, z_i) = (W_{S^i} x_i - y_i)^2 = (W_{S^i} x_i - W_S x_i)^2 = ((W_{S^i} - W_S) x_i)^2 \tag{5}$$

We have

**Theorem 1** *(very informal) For distributions $P_S$ for which a given regressor (or classifier) has an expected zero empirical error on the training set, $CV_{loo}$ stability in expectation is equivalent to expected error of the regressor (or classifier).*

*Remark*

The same result can be obtained for *quasi-ERM*, which selects an almost minimizer of the empirical risk, in the limit of $n \to \infty$ by using the *almost positivity* property of quasi-ERM.

In the following we consider the expected error term in $\text{CV}_{loo}$ stability, effectively assuming that the empirical error is negligeable.

# 3 Stability and Minimum Norm

I conjecture that asymptotically for $t \to \infty$ the minimum norm solutions have optimum stability among the zero-loss solutions provided by ERM in the overparametrized case. I do not know how to prove this in full generality. It has been proven for kernel machines. I will state it as a conjecturefor deep networks. The conjecture is

**Conjecture 2** *The solutions for $f_S$ satisfying $V(f_S, z_i) = 0, \quad \forall i$ that are minimum norm are the most stable.*

For later use, I recall the following result, linking minimum norm and maximum margin in the case of classification (see [15]):

**Lemma 3**

*The maximum margin problem*

$$\max_{W_K, \cdots, W_1} \min_n y_n f(W; x_n), \quad \textit{subj. to} \quad \|W_k\| = 1, \quad \forall k. \tag{6}$$

*is equivalent to*

$$\min_{W_k} \frac{1}{2} \|W_k\|^2, \ \textit{subj. to } y_n f(W; x_n) \geq 1, \quad \forall k, \ n = 1, \ldots, N. \tag{7}$$

The conjecture presents a puzzle. Consider regression under the square loss: in general with large overparametrization we expect several solutions with the same minimum norm. Will they have the same stability and thus the same expected errror? Or is an additional property needed, beyond the norm, to characterize the most stable solution?

### 3.1 Linear Regressors

The first argument is about linear functions $f_S(z_i) = W_S x_i$. Fitting the training set provides the set of $n$ equations

$$W_S X - Y = 0 \tag{8}$$

Assume $W_S \in \mathbb{R}^{1,d}$, $X \in \mathbb{R}^{d,n}$ and $Y \in \mathbb{R}^{1,n}$ with $n < d$. Then there are an infinite number of solutions for $W_S$ given by $W_S = YX^\dagger + (I - XX^\dagger)z$ where $z$ is any vector. The solution of minimum norm is $W_S = YX^\dagger$.

Let us explain the intuition that the minimum norm solution is the most stable. The minimum norm solution among all the infinite solutions is $W_S = YX^\dagger$. In the case in which $S$ is perturbed by deleting one data point we expect the change $\Delta X$ in $X$ to be small and decreasing with $n$. Consider $W_{S^i} = (Y + \Delta Y)(X + \Delta X)^\dagger$. Suppose $X$ is a $d, n$ matrix with $n < d$. Then $X^\dagger = (X^T X)^{-1} X^T$ and $(X + \Delta X)^\dagger = ((X + \Delta X)^T (X + \Delta X))^{-1} (X + \Delta X)^T$ Let us assume that $||\Delta X||$ is small and $||(X^T X)^{-1}||$ is large. Let us call $X^T X = A$ and $\Delta X = \Delta$.

Then $(X + \Delta)^\dagger \approx (A + X\Delta^T + (\Delta X^T)^{-1}(X + \Delta)^T$. Consider $(A + X\Delta)^T + \Delta X^T)^{-1} \approx A^{-1} - A^{-1}(X^T \Delta X + \Delta X^T X)A^{-1}$. Thus $(X + \Delta)^\dagger \approx [A^{-1} - A^{-1}(X^T \Delta + \Delta^T X)A^{-1}][(X + \Delta)^T]$. Putting things together and inspecting the various terms shows that $W_{S^i} = W_S + D$ where $D$ are terms that all contain the factor $A^{-1}$ and delta factors in either $X$ or $Y$ or both. The conclusion is $||W_{S^i} - W_S|| \approx ||(XX^T)^{-1}(\Delta X + \Delta Y)||$. In other words stability depends on $||(XX^T)^{-1}||$ an therefore on the norm $||W||$. This proof sketch should be cleaned up to show that *the minimum norm solution is the most stable solution and viceversa.* An obvious observation is that the same argument about the behavior of $||X^\dagger||$ in [16] can be used here. It shows that for random inputs $X$, $\mathrm{CV}_{loo}$ stability is expected to exhibit a double-descent curve implying a double-descent curve for the expected errror.

### 3.2 Deep Networks

Let us first introduce some notation. We define a deep network with $K$ layers with the usual coordinate-wise scalar activation functions $\sigma(z) : \quad \mathbf{R} \to \mathbf{R}$ as the set of functions $f(W; x) = \sigma(W^K \sigma(W^{K-1} \cdots \sigma(W^1 x)))$, where the input is $x \in \mathbf{R}^d$, the weights are given by the matrices $W^k$, one per layer, with matching dimensions. There are no bias terms: the bias is instantiated in the input layer by one of the input dimensions being a constant. We consider the case in which $f$ takes scalar values, implying that the last layer matrix $W^K$ is has size 1 x $h_{K-1}$, where $h_k$ denotes the size of layer $k$. The weights of hidden layer $k$ has size $h_k \times h_{k-1}$. In the case of of binary classification which we consider here the labels are $y \in \{-1, 1\}$. The activation function is the ReLU activation. For the network, homogeneity of the ReLU implies $f(W; x) = \prod_{k=1}^K \rho_k f(V_1, \cdots, V_K; x)$, where $W_k = \rho_k V_k$ with the matrix norm $||V_k||_p = 1$ and $||W_k|| = \rho_k$.

There are several ways to show that changes in the weights due to small changes in the training set will be proportional to the norm of the weights. A simple observation goes as follows. In a deep net, the product of the norms in a $K$-layer networks is $\rho_1 \cdots \rho_K$. Since we know that if

the $\rho_k$ start equal then they grow at the same rate under gradient descent and thus remain equal (see [15]), we assume that the total norm of the network is $\rho^K$ (the argument is valid even if the $\rho_k$ are different). Assume now that the weights of each layer are perturbed because of a change, such as leave-one-out, in the training set . Then the overall norm will change as

$$\rho^K \to K\rho^{K-1}\Delta\rho, \tag{9}$$

implying that for $V(f,z) = c_\gamma(f(x),y)$ as defined in section 4.2.2 of [17]

$$V(f_{S^i}(x_i) - f_S(x_i)) \leq \frac{1}{\gamma}||f_{S^i}(x_i) - f_S(x_i)||||x|| \leq \frac{1}{\gamma}\rho^{K-1}(\rho - \Delta\rho) \tag{10}$$

Thus minimizing the norm $\rho$ (for a fixed margin) minimizes a bound on $E_S|Vf_S^i(x_i) - f_S(x_i))|$, that is on $\mathrm{CV}_{loo}$ stability. The same argument is valid for other loss functions such as the square losss.

### 3.3 A General Approach?

A possibly more general approach to establish that stable solutions are minimum norm and viceversa may rely on the *implict function theorem* or on the more powerful *constant rank theorem*. The observation is that fitting the training set corresponds to the equation

$$F(X,Y,W) = 0 \tag{11}$$

where $X^*, Y^*$ is the training set, $W$ is the set of weights and $F(X,Y,W)$ is a set of $n$ equations for each of the data points (columns of $X$ and $Y$). Under assumptions of differentiability of $F$, the interpolating or separating property defines a mapping $W(X,Y)$ in the neighborhood of the solution $X^*, Y^*, W^*$ such that $F(X,Y,W(X,Y)) = 0$ in that neighborhood. Furthermore $\frac{\partial W}{\partial X}$ may be computed in terms of the Jacobian of $F$ and other derivatives. In the case of $F(X,Y,W) = WX - Y$, this approach would then provide $\Delta W(X) =\approx \frac{\partial W}{\partial X}\Delta X \approx X^\dagger \Delta X$. Thus

**Conjecture 4** *(very informal) Using the implicit function theorem, $CV_{loo}$ stability for kernel regressors+classifiers and for deep nets can be bounded by the norm of the weights. Thus minimum norm solutions (xslocally) optimize stability.*

### 3.4 Hard margin SVM

In the case of hard margin linear SVM it is not clear in terms of the classical theory (there are separate arguments, such as the perceptron learning theorem) why one should select the maximum margin solution among all the separating hyperplanes. Our approach provides an answer: one must choose the most stable solutions in order to minimize the expected error, and the most stable solution in the case of hard margin linear SVM is the minimum norm one for margin equal to 1 (which is equivalent to the maximum margin solution, see section in [15] on maximum margin and minimum norm).

### 3.5 Square loss in deep networks is biased toward minimum norm solutions because of implicit dynamical regularization during GD

- We set $f(x) = \rho f_V(x)$ with $\rho, V, f_V$ defined as in [15];
- we assume $||x|| = 1$ implying $||f_V(x)|| \leq 1$ at convergence;
- observe that if margin of $f_v$ on $S = x_n, y_n, n = 1, \cdots, N$ is 1, then $y_n f_V(x_n) = 1, \forall n$, that is all training points are support vector with maximum margin value;
- **Lemma 5** *[15] The maximizer of the margin when $||V_k|| = 1$ is the minimum norm solution when the margin is $\geq 1$.*

Let us consider the dynamical system induced by GD on a deep net with RELUs. We change variables by using $W_k = \rho_k V_k$, $||V_k|| = 1$. From now on, I will use $f$ for $f_v$. Gradient descent on $L = \mathcal{L} + \lambda \sum_k V_k^2 = \sum_n (\rho f_n - y_n)^2 + \lambda \sum_k V_k^2$ yields the dynamical system

$$\dot{\rho_k} = -\frac{\partial L}{\partial \rho_k} = -2 \sum_n (\rho_k^L f_n - y_n) f_n \rho_k^{L-1} = -2\rho_k^{L-1} [\sum_n \rho_k^L (f_n)^2 - \sum_n f_n y_n] \tag{12}$$

and

$$\dot{V_k} = -\frac{\partial L}{\partial V_k} = -2 \sum_n (\rho f_n - y_n) \rho \frac{\partial f_n}{\partial V_k} - 2\lambda V_k \tag{13}$$

Following Shai [18] we use Equation 12 to derive the dynamics of $\rho = \rho_k^L$ in terms of $\dot{\rho} = \sum_k \frac{\partial \rho}{\partial \rho_k} \dot{\rho_k}$

Thus

$$\dot{\rho} = 2L\rho^{\frac{2L-2}{L}} [-\sum_n \rho (f_n)^2 + \sum_n f_n y_n] \tag{14}$$

Because of the constraint imposed via Lagrange multipliers $||V_k||^2 = 1$, then $V_k^T \dot{V_k} = 0$, which gives $\lambda = -\sum_n (-\rho^2 f_n^2 + \rho y_n f_n)$. Thus

$$\dot{V_k} = 2 \sum_n [(\rho f_n - y_n) \rho (-\frac{\partial f_n}{\partial V_k}) - 2V_k \rho f_n (\rho f_n - y_n)] \tag{15}$$

that is

$$\dot{V_k} = 2\rho \sum_n [(\rho f_n - y_n)(-V_k f_n - \frac{\partial f_n}{\partial V_k})] \tag{16}$$

Observe that $\dot{\rho} = 0$ if $y_n f_n = 1$ and $\rho_k = 1$; in general the equilibrium value for $\dot{\rho_k} = 0$ is $\rho = \frac{\sum_n y_n f_n}{\sum_n f_n^2}$.

Values $y_n f_n = 1, \rho = 1$ are stationary points of the dynamics of $V_k$ given by $\dot{V_k} = 0$: they are minimizers with zero square loss. In general the stationary points of $V_k$ are given by (using $\ell_n = \rho f_n - y_n$ and assuming $\ell_n \neq 0$) by

$$V_k = -(\sum_n \frac{\partial f_n}{\partial V_k}(y_n - \rho f_n))(\sum f_n(y_n - \rho f_n))^{-1} = -\frac{\sum_n \frac{\partial f}{\partial V_k}\ell_n}{\sum_n f_n \ell_n}. \quad (17)$$

The question is why the dynamics should be biased towards $\min \rho$ provided that $\rho f_i \geq 1, \quad \forall i$. Notice that the lowest possible value of $\rho$ is $\rho = 1$ which can be achieved if $y_n f_n$ is either $= 1$ or $= 0$ – which corresponds to *maximum sparsity wrt training set*. Notice that the constrained problem *minimize* $L = \sum(\rho f(x_i) - y_i)^2$ *s.t.* $\rho = const$ has the same form independently of whether $const = 1$ or $< 1$ (is this true? $\lambda$ is different). Also notice that this is the Ivanov version of Tikhonov regularization minimizing the norm $\rho = \rho_k^L$. I conjecture that if the initial conditions are $\rho_{t=0} \approx 0$ and at least some of the $y_n f_n < max_V y_n f_n = 1$ then $\rho(t)$ grows but very slowly for a longish time until it grows very quickly to its asymptotic value. The dynamics of Equation 14 is that (for "largish" $K$) the smaller $\rho_{t=0}$ is, the longer it takes to $\rho$ to grow (a similar dynamics in a different context is show in Figure 2 in [18]).

The intuition is that $\rho$ is constrained to be very small and roughly constant value for a long time during GD iterations (until $T$: $T$ is longer with more layers and longer with smaller initialization). At around $T$, $\rho$ will grow very quickly to a value which depends on $\frac{\sum y_i f(x_i)}{\sum f^2(x_i)}$. The idea is that this dynamics (from $t = 0$ to $t = T$) is similar to minimizing the square loss under the constraint of a small constant $\rho$, which is itself equivalent to Tikhonov regularization (which minimizes the norm).

This is similar (???) to regularization that penalizes $||f||$ for a finite but long time and then released.

### 3.6 Gradient Descent (GD) and Selection of Minimum Norm Solutions

Until now I have discussed ERM, without discussing the optimization algorithm used for minimization. The summary is that in order to ensure good expected error for interpolating regressors, it is necessary to select the most stable solutions and to do that one needs to select the minimum norm solutions among all the infinite solutions that achieve zero empirical loss. So ERM is not enough by itself in the overparametrized case. However, it turns out that if GD is used to perform ERM, GD will select among the empirical minimizers the ones with minimum norm both in the case of kernel regression ([13]) and of deep networks.

## 4 Caveats

In summary, the two main claims of this paper are 1) that minimizing $\text{CV}_{loo}$ stability minimizes the expected error and 2) choosing the minimum norm solutions among all the solutions with zero empirical error minimizes stability (locally). This latter claim is proven so far for kernel machines; an outline of a possible proof was given in this paper for deep networks under exponential-type losses.

It is now important to derive more formal bounds for both the case of kernel regressors and the case of deep networks. Two papers in preparation will [19, 20] describe those results.

## 5 Conclusions

In summary, optimization of $CV_{loo}$-type stability minimizes for $n \to \infty$ the expected error in both the classical and the modern regime of ERM. It is thus *a sufficient condition* for predictivity in ERM (but probably beyond ERM, see [1]).

In the classical regime, stability implies generalization and consistency. In the modern regime, stability probably explains the double descent curve in kernel interpolants [19] and why maximum margin solutions in deep networks trained under exponential-type losses may minimize expected error (this does not mean they are globally optimal), see [20].

Conditions for learnability and stability in learning theory may have deep, almost philosophical, implications: as remarked by V. Vapnik, they can be regarded as equivalent conditions that guarantee any scientific theory to be predictive and therefore "scientific". The condition coming from classical learning theory corresponds to choosing the theory from a fixed “small” set of theories that best fit the data. The condition prescribed by the modern theory corresponds to choosing a theory from a “large” hypothesis set (that can even increase before new data arrive) that fits the data *and* is simplest (Occam razor, Einstein). These two conditions can be summarized and unified by the principle of selecting the most stable theory — the one that most of the time changes the least if data are perturbed or when new data arrive. Thus Thomas Kuhn scientific revolutions are allowed, as long as they do not happen too often!

**Acknowledgments** We thank Gil Kur, Andy Banbuski, and especially Silvia Villa and Lorenzo Rosasco. This material is based upon work supported by the Center for Minds, Brains and Machines (CBMM), funded by NSF STC award CCF-1231216, and part by C-BRIC, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA. This research was also sponsored by grants from the National Science Foundation (NSF-0640097, NSF-0827427), and AFSOR-THRL (FA8650-05-C-7262) and was also supported by the Institute for Information and Communications Technology Promotion (IITP) grant funded by the Korea government(MSIT) (No.2019-0-00079, Department of Artificial Intelligence(Korea University))..

**Correspondence** Correspondence and requests for materials should be addressed to T.Poggio (email: tp@ai.mit.edu).

## References

[1] T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi. General conditions for predictivity in learning theory. *Nature*, 428:419–422, March 2004.

[2] Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability

is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1):161–193, 2006.

[3] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *J. Mach. Learn. Res.*, 11:2635–2670, December 2010.

[4] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *CoRR*, abs/1611.03530, 2016.

[5] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[6] Madhu S. Advani and Andrew M. Saxe. High-dimensional dynamics of generalization error in neural networks. *arXiv e-prints*, page arXiv:1710.03667, Oct 2017.

[7] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *CoRR*, abs/1903.07571, 2019.

[8] M. Belkin, S. Ma, and S. Mandal. To understand deep learning we need to understand kernel learning. *ArXiv e-prints*, Feb 2018.

[9] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv e-prints*, page arXiv:1908.05355, Aug 2019.

[10] Alexander Rakhlin and Xiyu Zhai. Consistency of Interpolation with Laplace Kernels is a High-Dimensional Phenomenon. *arXiv e-prints*, page arXiv:1812.11167, Dec 2018.

[11] Tengyuan Liang and Alexander Rakhlin. Just Interpolate: Kernel "Ridgeless" Regression Can Generalize. *arXiv e-prints*, page arXiv:1808.00387, Aug 2018.

[12] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in High-Dimensional Ridgeless Least Squares Interpolation. *arXiv e-prints*, page arXiv:1903.08560, Mar 2019.

[13] Lorenzo Rosasco and Silvia Villa. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems*, pages 1630–1638, 2015.

[14] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal Machine Learning Research*, 2001.

[15] A. Banburski, Q. Liao, B. Miranda, T. Poggio, L. Rosasco, B. Liang, and J. Hidary. Theory of deep learning III: Dynamics and generalization in deep networks. *CBMM Memo No. 090*, 2019.

[16] T. Poggio, G. Kur, and A. Banburski. Double descent in the condition number. *CBMM memo 102*, 2019.

[17] O. Bousquet and A. Elisseeff. Algorithmic stability and generalization performance. In *Neural Information Processing Systems 14*, Denver, CO, 2000.

[18] Daniel Gissin, Shai Shalev-Shwartz, and Amit Daniely. The Implicit Bias of Depth: How Incremental Learning Drives Generalization. *arXiv e-prints*, page arXiv:1909.12051, September 2019.

[19] Lorenzo Rosasco, Gil Kur, and Tomaso Poggio. Stability of kernel regression in the modern regime. *in preparation*, 2020.

[20] Tomaso Poggio and et al. Stability of deep networks. *in preparation*, 2020.